Abstract

An system for segmenting strings into component parts for use with a database management system. A reference table of string records are segmented into multiple substrings corresponding to database attributes. The substrings within an attribute are analyzed to provide a state model that assumes a beginning, a middle and an ending token topology for that attribute. A null token takes into account an empty attribute component and copying of states allows for erroneous token insertions and misordering. Once the model is created from the clean data, the process breaks or parses an input record into a sequence of tokens. The process then determines a most probable segmentation of the input record by comparing the tokens of the input record with a state models derived for attributes from the reference table.